3

Implicature

Recall from Chapter 1 that the logical connectives frequently are used to mean something beyond, or different from, their logical meaning. For example, *or* is logically inclusive; $p \vee q$ ('p or q') is true not only when one of the two propositions is true, but also when they're both true. But we often use it in an exclusive sense:

(1) I'll rewrite this chapter or I'll delete it.

This is typically taken to mean that I'll either rewrite the chapter or delete it, but not both. That 'but not both' aspect of the meaning goes beyond the logical meaning of the connective. Similarly, $p \wedge q$ is true just in case both conjuncts are true, but *and* is frequently used with a meaning of ordering or causation that is not part of its logical meaning:

(2) a. This morning I had a cup of coffee and went out for a walk.
b. They had spent two afternoons at the creek, they said they were going in naked and I couldn't come…
(H. Lee 1960, *To Kill a Mockingbird*)

In both cases in (2), the sentence is true if both of the component propositions are true—that is, if the speaker in (2a) both went out for a walk and had a cup of coffee at some point in the morning, and if the swimmers in (2b) said they were going in naked and also said that the speaker couldn't come. But of course what's probably meant in (2a) is that the events happened in the order stated: that the speaker first had a cup of coffee and then went out for a walk; and what's meant in (2b) is

that the reason that the speaker couldn't come was that the others were going in naked. As a final example, note that the use of the conditional often has a biconditional interpretation:

(3) If you behave at the store, we'll stop for ice cream afterward.

The utterance in (3) logically says nothing about what will happen if you don't behave, but the intended meaning of such a statement is usually taken to be 'if you behave at the store, we'll stop for ice cream afterward, but if you don't behave, we won't'.

So where are all these extra bits of meaning coming from?

The Cooperative Principle

H. P. Grice (1975) attributed the difference to a concept that is both sweeping in its coverage and elegant in its simplicity: He said essentially that much of what we understand a speaker to have meant is based on our assumption that they are being cooperative. He termed this the Cooperative Principle (CP). Although the CP boils down to, essentially, 'be cooperative', the full CP is slightly less elegant but a lot more precise:

The Cooperative Principle: "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged." (Grice 1975)

You might not consider this an especially cooperative way of phrasing what boils down to 'be cooperative', but a lot of those extra words are there to emphasize the importance of context: Your contribution should be what's required **in context**: when it occurs, in the conversation in which it occurs, for the accepted purposes of the people among whom it's occurring.

One thing to notice right off the bat is that the CP is phrased prescriptively, but it's actually descriptive: It says, essentially, 'do this'—but what it means is 'people consistently do this'. After all, nobody has to tell you to (for example) make your utterances relevant to the ongoing conversation. Nobody ever slaps their forehead and says, "Oh, when you asked where the bathroom is, you wanted a *relevant* answer!

I had no idea!" This is the case for all of the 'rules' linguists talk about; although we may phrase them as directives, they are always descriptions of the way people actually behave. Remember, prescriptive rules like 'don't split an infinitive' exist precisely because people DO split infinitives, all the time, and someone wants us to stop; descriptive rules like 'be relevant' are the ones we follow without ever having to be told, and those are the rules linguists are concerned with.

So in the CP, Grice has formulated a rule describing how people behave in conversation—and in brief, it says that we behave cooperatively. In order to see how this helps to solve his problem (which, remember, was that there's a big difference between sentence meaning and speaker meaning, and more specifically between the logical connectives and their natural-language use), we'll need to look more closely at the CP and its associated 'maxims'. Grice placed his maxims into four categories; over time these four categories have themselves come to be known as the four maxims of the CP (a slight blip in the terminological history, but we'll stick with current usage). As Grice phrased them, they are:

Quantity:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

Quality: Try to make your contribution one that is true.

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Relation:

1. Be relevant.

Manner: Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

In addition, there are several ways you can use these maxims: You can **fulfill** them, you can **violate** them, you can **flout** them (violating them in an exaggerated and obvious way), or, finally, you can **opt out** altogether. These actions, in turn, can generate an **implicature**, i.e., they can invite the hearer to infer that the speaker meant more than they semantically said. If I tell you *I'm thirsty* but what I really mean is 'please bring me a drink', what I've **said** is that I'm thirsty, and what I've **implicated** is a request for you to bring me a drink. And if all goes well, that will be what you **infer**, too. That's a case where I've fulfilled the maxims—but I can get the same effect by flouting the maxim of Quality, by saying something so exaggeratedly false that you know I'm not stating the literal truth; for example, I can say *I'm dying of thirst*. I'm not really dying, and it's perfectly obvious to you that I'm not dying, but you're still likely to infer that I'm asking for a drink.

Violating a maxim can generate an implicature too, but in this case it's an implicature intended to mislead the hearer. And opting out is just what it sounds like; I'm trying to hold a conversation, but you're reading the newspaper, playing the guitar, wandering off, or in some other way pointedly choosing not to participate. In a courtroom, pleading the Fifth Amendment is a way of opting out (although doing so might give rise to its own implicature—e.g., in the courtroom case, that the speaker is guilty).

Notice that the ability to generate implicatures is a huge advantage to both speakers and hearers: First, speakers save an enormous amount of time by not having to spell out every single aspect of their meaning (e.g., for (2a), *This morning I had a cup of coffee and then after that I went out for a walk*), especially the meanings that are slightly less pleasant or polite, which the speaker may want to leave 'off-record'. And it also allows the speaker to convey other meanings without being committed to them: In (3), by saying *If you behave at the store, we'll stop for ice cream afterward*, the speaker implicates 'if you don't behave, we won't stop for ice cream', but is not committed to having said so explicitly, and thus leaves the door open for going ahead and buying ice cream anyway after the hearers have behaved badly. (And as we'll see later, implicature also helps speakers and hearers negotiate the tension between saying as much as, but no more than, necessary.)

We'll consider each of the maxims in turn, providing examples of ways they can be used in order to generate an implicature. And regarding terminology, remember: Speakers implicate; hearers infer. Implicating is very different from implying (which in linguistics means something much like entailing, so it's safest just to remember that 'imply' is not used in talking about pragmatics). And when what a hearer infers is different from what a speaker implicated, the result is miscommunication.

The Maxim of Quantity

The Maxim of Quantity has two parts:

1. Make your contribution as informative as is required (for the current purposes of the exchange).
2. Do not make your contribution more informative than is required.

This presents a nice tension: Say enough, but don't say too much. Most of the Quantity-based implicatures discussed in the literature are based on the first submaxim (in part because of an interesting relationship between the second submaxim and the maxim of Relation, which we'll discuss shortly). Because a speaker is assumed to be saying as much as is required, the hearer will assume that the speaker could not have said more without being uncooperative in some other way. For example, let's take the common phenomenon known as **scalar implicature** (Horn 1972), exemplified in (4):

(4) a. I've washed most of the windows.
b. Jordan ate half of the pizza.
c. There are several birds at the feeder.

In (4a), the hearer is likely to infer that the speaker hasn't washed all of the windows, since if they had, they should have said so (in order to count as 'saying enough'). Likewise, the speaker in (4b) implicates that Jordan didn't eat any more than half of the pizza, and the speaker in (4c) implicates that there aren't, say, dozens of birds at the feeder—that is, that there are no more than several. In each case, there's an implicit

scale being invoked, and the speaker's choice of a value on the scale implicates that no higher value holds, since if it did, they should have said so. So in (4c), we can imagine a scale of amounts in which *several* is a higher value than, say, *a couple*, but a lower value than *dozens*. By selecting the value *several*, the speaker implicates (among other things) 'not dozens'.

It's important to note that the scales in question are 'Horn scales' (Horn 1972), which are ranked from semantically stronger to semantically weaker expressions. For example, if one expression entails another (and the entailment doesn't go both ways), the entailing expression is stronger than the entailed expression—so, since to wash *all of the windows* entails washing *most of the windows*, that means that *all of the windows* is a semantically stronger expression, and is higher on the scale. In turn, choosing to say *most of the windows* implicates that the stronger expression *all of the windows* would not have been appropriate, and hence that **only** most of the windows, but not all, were washed. (See Hirschberg 1991 for discussion of other types of scales and their effects on implicature.)

For another Quantity-based implicature, consider (5):

(5) "This is your mother," said Dorothea, who had turned to examine the group of miniatures. "It is like the tiny one you brought me; only, I should think, a better portrait. And this one opposite, who is this?"

"Her elder sister. They were, like you and your sister, the only two children of their parents, who hang above them, you see."

"The sister is pretty," said Celia, implying that she thought less favorably of Mr. Casaubon's mother.

(George Eliot, 1871, *Middlemarch*)

Setting aside the use of the word *imply* in the last sentence (again, a linguist would say *implicate*), why is Celia taken to mean that the mother isn't as pretty?

The answer, of course, lies in the Maxim of Quantity. If Celia thought both women were pretty, she should have said so, since both of them are under discussion. By stating only that the sister is pretty, she implicates that she is not in a position to say that both of them are, and hence that

the mother is not. In essence, there's a scale on which *one sister is pretty* is a lower value than *both sisters are pretty*, so to affirm the lower value implicates a denial of the higher value.

All of these instances involve cases in which the maxim is fulfilled. What if it is violated? To violate the maxim would be to simply not give enough information, or to give too much, and it generally leads to an intended but inappropriate inference. Suppose you're thinking of hiring my niece Jane Doe, and you ask me whether she'd be a good employee. I tell you truthfully that she's hard-working, smart, ambitious, and organized—but I fail to mention that she was fired from her last job for stealing money from the cash register and screaming at the customers. Surely I have not said enough; I've violated the Maxim of Quantity. And in doing so, I've given my reader the incorrect impression that Jane would be just dandy as an employee.

We saw a real-world violation of Quantity in Chapter 2, example (11), in which President Clinton said *It depends on what the meaning of the word 'is' is*. The statement in question is 'there is absolutely no sex of any kind in any manner, shape or form, with President Clinton', and Clinton is arguing that that's a truthful statement because the word *is* is present-tense and there was no sex going on at the time of the statement. So here we have a semantically truthful answer. But it's an uncooperative answer, in that it violates the Maxim of Quantity. Clearly what the questioner wants to know is whether any sex ever took place; because Clinton knows this, he's saying less than is required—so he's violating the maxim, and he knows perfectly well that by committing this quiet violation he will lead his hearers to assume that he has said as much as is required, implicating that he has never had a sexual relationship with Lewinsky. In this case a violation of the maxim is successfully used to generate a misleading implicature.

So we've seen implicatures generated by speakers fulfilling the maxim and by speakers violating it. What about flouting? This is where things get especially interesting. In flouting a maxim, the speaker violates it so egregiously that the hearer can't help but notice. So although they're not really fulfilling the maxim, the speaker is still behaving cooperatively, and certainly isn't trying to mislead the hearer. Take, for example, the case of 'damning with faint praise': If I've set you up on a blind date with my friend and you ask what she's like, I had better say more than *she's nice*. If you

ask me for a letter of recommendation, that letter had better comment on more than your handwriting. And if you ask a friend how they like their new boss, they'd better comment on more than his cufflinks. In any of these cases, the extent to which the response falls short of what was expected will lead the hearer to infer that a negative assessment was intended.

A beautiful real-world instance of a flouting of Quantity was provided by Henry Kissinger when *Time* magazine (in compiling its list of the hundred most influential people of 2017) asked him to write a brief piece on Jared Kushner. In general, these pieces are tributes or encomiums—that is, high praise. Here's what Kissinger wrote:

(6) Transitioning the presidency between parties is one of the most complex undertakings in American politics. The change triggers an upheaval in the intangible mechanisms by which Washington runs: an incoming President is likely to be less familiar with formal structures, and the greater that gap, the heavier the responsibility of those advisers who are asked to fill it.

This space has been traversed for nearly four months by Jared Kushner, whom I first met about 18 months ago, when he introduced himself after a foreign policy lecture I had given. We have sporadically exchanged views since. As part of the Trump family, Jared is familiar with the intangibles of the President. As a graduate of Harvard and NYU, he has a broad education; as a businessman, a knowledge of administration. All this should help him make a success of his daunting role flying close to the sun.

(http://time.com/collection/2017-time-100/4742700/jared-kushner/)

This is, to my mind, a wonderful piece of damning with faint praise. It's generally positive, yet consider it in light of the Maxim of Quantity: It leaves out precisely what it should have included, which is something specific about Kushner's qualifications or achievements. The first paragraph essentially states that Kushner's task is a difficult one. The first half of the second paragraph says that Kissinger knows Kushner. Only in the

last three sentences does he state anything about Kushner's qualifications, and what he says is mild compared to what is expected: He notes that Kushner is 'familiar with the intangibles of the President', that he has 'a broad education', and that he has 'a knowledge of administration'. And he closes by invoking the myth of Icarus, whose flight 'close to the sun', as we all remember, ended in disaster. In light of the contextual expectation of high praise, this falls stunningly short. And not surprisingly, many readers took it as implicating a negative overall assessment. (See Blake 2017 for a tidy analysis.)

The Maxim of Quality

The Maxim of Quality states:

Try to make your contribution one that is true.

1. Do not say what you believe to be false.
2. Do not say that for which you lack adequate evidence.

Interestingly, it doesn't say 'Make your contribution one that is true', but rather 'TRY to make your contribution one that is true'. This is a nice implicit acknowledgment that we can't possibly know for certain what is and isn't true, and it's reinforced by the submaxims. How do you try to make your contribution true? Well, by not saying what you believe to be false, and by not saying that for which you lack evidence.

I think it's fair to say that most of the time we obey the Maxim of Quality, by saying things that we do believe to be true and that we do have evidence for. A quiet violation of the first submaxim of Quality is a straightforward lie: If you say something you believe to be false, you'll (probably intentionally) mislead your hearer. And while that's often a bad thing, it's worth remembering those 'little white lies' that help to keep our relationships running smoothly:

(7) A: How do you like my new dress?
B: It's great!

Needless to say, if B doesn't think the dress is great at all, then they've lied, but that's not necessarily a bad thing. You might object that B could have avoided the lie by simply quietly violating the Maxim of Quantity instead, and not saying as much as is called for:

(8) A: How do you like my new dress?
B: I like the color!

If B does like the color but hates the style, avoiding the problem by failing to answer the specific question at hand and instead answering a closely related question may solve the problem. Unfortunately, people are pretty good at picking up on this sort of equivocation and will frequently draw a scalar Quantity-based inference such as (in this case) 'B has said they like the color of my dress; a higher value on the scale of my dress's properties would have included the color, the style, the cut, the fit, and how it suits me; because B has chosen to state only that they like the color, I can infer that B does not like these other properties'. (In short, beware: Gricean implicature can get you out of a sticky situation or can make the situation a lot stickier.)

So is a violation of Quality the exact same thing as a lie? As it happens, people differ on this. Remember Prototype Theory from Chapter 1? The general idea was that a word like *sandwich* couldn't be defined in terms of a strict set of features, because there are 'fuzzy' cases that people can't quite agree on (like hot dogs). Coleman and Kay (1981) argue that the meaning of the word *lie* is another instance of Prototype Theory at work. In their view, a prototypical lie has three properties: (1) it is false, (2) the speaker believes it is false, and (3) the speaker intends to deceive the addressee. The more of these features an utterance has, Coleman and Kay found, the more likely a subject is to consider that utterance to be a lie. And their most interesting finding (at least with respect to the CP) is that the most important factor in determining whether an utterance was considered a lie was #2: The speaker believes it to be false. And that, of course, is precisely what's forbidden by the first submaxim of Quality: 'Do not say what you believe to be false'. In short, there are utterances that are 'kind of' lies or 'just barely' lies (e.g., you say something false when you believe it's true, or you say something true but with the

intention to deceive, as with the Clinton 'meaning of *is*' example), but the clearest lies are instances when the speaker directly violates the Maxim of Quality by saying something they believe to be false.

Meanwhile, what constitutes a flouting of Quality? That is, when would you want to say something so egregiously false that you want the hearer to realize you're saying something false? It sounds unlikely at first, but when you think about it for a moment you'll realize that we do it all the time. Consider the examples in (9):

(9) a. It is not an exaggeration to say that burgers are America.
(*Costco Connection*, June 2018)
b. I've got a ton of onions on this burger.
c. I'm parked in front of the burger joint.

Semantically speaking, in a typical scenario all three of these utterances are likely to be false. Burgers are not, in fact, America; America is not made of ground beef. So while in (9a) the writer is correct to say it's not an exaggeration, it is nonetheless a metaphor—and a flouting of Quality. And other floutings of Quality are indeed exaggerations; for example, nobody could fit a literal ton of onions on a standard-sized hamburger, so (9b) is a case of exaggeration, aka hyperbole. And finally—and perhaps a bit more subtly—it's not the speaker in (9c) who is parked in front of the burger joint, but rather the speaker's car. Similarly, consider (10):

(10) Izzy, who had been playing violin since she was four, and had been assigned second chair even though she was a freshman, should have had nothing to fear. "You'll be fine," the cello had told her, eyeing Izzy's frizzy golden hair—the dandelion fro, Lexie liked to call it.
(Ng 2017, *Little Fires Everywhere*)

Here, it's safe to assume that the cello itself hasn't spoken up; rather, the writer is referring to the cellist by the use of the noun phrase *the cello*.

Finally, cases of sarcasm or irony (e.g., *He's a real Einstein* or *Another beautiful February day in Chicago*) are floutings of Quality, with the hearer expected to recognize that the speaker is saying something they obviously don't believe to be true.

In all of these cases, the speaker has violated the Maxim of Quality so egregiously that it's assumed the hearer will notice; if the reader in (10) wasn't expected to notice the flouting, the novel would have been seriously derailed (consider the reader reacting to (10) with a shocked, "The cello just spoke!"). The hearer, believing the speaker is being cooperative, will search for an interpretation under which the utterance makes sense.

The Maxim of Relation

This one is very brief, but behind its brevity lies a good deal of insight:

Be relevant.

That's it; no expansion or submaxims. And of course most of the time we are indeed being relevant. And if at first our utterance doesn't seem relevant, the hearer's overarching belief that the speaker is trying to be cooperative will lead them to search for some interpretation on which our utterance is in fact relevant. For example, consider this account of an interview with Stormy Daniels, who claimed to have had an affair with Donald Trump:

(11) Daniels said she was in a parking lot preparing to go into a fitness class, and was pulling her infant daughter's car seat and diaper bag out of her vehicle.

"And a guy walked up on me and said to me, 'Leave Trump alone. Forget the story'," Daniels said. "And then he leaned around and looked at my daughter and said, 'That's a beautiful little girl. It'd be a shame if something happened to her mom'. And then he was gone."

(https://www.cnn.com/2018/03/25/politics/60-minutes-stormy-daniels-interview-main/index.html)

Why is this example so disturbing? It's because like Daniels, we take the statement *it'd be a shame if something happened to her mom* to be a

threat. On the face of it, of course, it's merely a non sequitur; it seems to be irrelevant to the primary topic, which is the Trump story. In order to interpret it as relevant, we supply the missing link: If you don't drop the story, something might happen to her mom. Semantically, it's an obvious truth: It would indeed be a shame. But pragmatically, we read it as a threat—and that reading depends on our assumption that the speaker is being relevant.

Needless to say, the Maxim of Relation can also be flouted in order to generate an implicature, as in (12):

(12) A: Did you see the ridiculous hat Chris was wearing?
B: Um, nice day we're having.

Here B's comment is so blatantly irrelevant that A can only assume that B is implicating a need to change the topic immediately—perhaps because Chris is standing within hearing range.

Violations of Relation are common, especially in political debates, in which a candidate will often seem to ignore the question being asked and will instead speak on some other topic. Here the goal isn't to generate an implicature, and the candidate doesn't especially want the audience to notice the switch; the goal is either to avoid an uncomfortable question or to spend the time talking about a preferred topic. But violations can be used purposely to mislead the hearer as well, as in (13):

(13) A: Did Frank enjoy his visit?
B: I hope so. All our vodka is gone.

Now, suppose B actually drank all the vodka but is trying to hide that fact. In that case, the comment *All our vodka is gone* is irrelevant, but because B knows that A will assume that the comment is relevant, A implicates that Frank is the culprit without (strictly speaking) lying. Again, the hearer's assumption that the speaker is being cooperative leads them to the most relevant interpretation; and the speaker, knowing this is what the hearer will do, can use that fact to lead the hearer to an interpretation that is intended but false.

The Maxim of Manner

The last maxim is a bit of a grab-bag. The Maxim of Manner states:

Be perspicuous.

1. Avoid obscurity of expression.
2. Avoid ambiguity.
3. Be brief (avoid unnecessary prolixity).
4. Be orderly.

So first, let's just roll our eyes at the extent to which the maxim appears to violate itself: There are plenty of ways to say 'be clear' besides 'be perspicuous' (which means 'be clear' but is less clear). And adding to 'be brief' the parenthetical expansion 'avoid unnecessary prolixity' is so unnecessarily prolix (i.e., nonbrief) that many assume it was intended as a joke.

But this maxim helps a lot with the problem we started out with, the original impetus for the CP, which is the difference between the semantic meaning of a logical connective and the range of meanings it is typically used for pragmatically. Consider the examples in (2), repeated here as (14):

(14) a. This morning I had a cup of coffee and went out for a walk.
b. They had spent two afternoons at the creek, they said they were going in naked and I couldn't come.

In (14a), the inference that the coffee preceded the walk is due to an implicature; as we've seen, this isn't part of the logical meaning of *and* (i.e., it's not part of the meaning of the corresponding logical operator, $\wedge$). Similarly, in (14b) the implicature that the nakedness is the reason why the speaker couldn't come is not part of the logical meaning of *and*. And by saying it's not part of the 'logical' meaning, I'm also saying that it's not part of the **semantic** meaning, if we're assuming a truth-conditional semantics, i.e., a semantics in which semantic meaning is the same as logical, truth-functional meaning.[1]

[1] Terminological hash: The truth-functional meaning of a logical operator is the function it performs on the expressions it connects, so for logical operators in a truth-conditional

So if the implicatures of ordering and causation in (14a)–(14b), respectively, aren't encoded in the meaning of *and*, where did they come from? Well, they came from the interaction of the Maxim of Manner and the utterance in context. Because Manner tells us to be orderly, we can assume that if two events happened in sequence, they'll be presented in the order in which they occurred. For events that didn't happen in sequence (as in *I had spaghetti and garlic bread for dinner*) or where the sequence is completely irrelevant (as in *What a day—I had three meetings and two conference calls*), the implicature won't arise. Or, as in (14b), a different implicature might be generated; in this case, there's no ordering to worry about, but the interaction between Relation (why mention nakedness if it's not relevant to what follows?) and Manner (if it's not relevant, it's unnecessarily 'prolix' to mention it) leads to the implicature that the first conjunct is the reason for the prohibition mentioned in the second.

This interaction among the maxims bears noting. You might have noticed that there's a certain amount of overlap between the second submaxim of Quantity—'Do not make your contribution more informative than is required'—and the Maxim of Relation: After all, what is it to be irrelevant, other than being more informative than is required? And there's overlap between both of them and the third submaxim of Manner—'be brief (avoid unnecessary prolixity)': Again, what is it to be unnecessarily prolix but to say more than is required, i.e., to say what is irrelevant? Hold that thought; we'll find that others have proposed alternative sets of maxims that take advantage of this tension between saying enough and not saying too much.

Meanwhile, however, what happens when we violate the Maxim of Manner? Grice gives the following example of a violation:

(15) A: Where does C live?
B: Somewhere in the South of France.

semantics, truth-functional meaning equals truth-conditional meaning equals semantic meaning, and logical connectives equal logical operators equal truth-functional operators equal truth-functional connectives (except for ¬ 'not', which is a logical operator but not a connective since it doesn't connect two things). If that makes you want to tear your hair out, ignore this whole footnote.

Here, A has been less clear (less perspicuous, more obscure) than one should, in light of the Maxim of Manner. What B is expected to infer is that A is not in a position to obey the maxim—i.e., that A doesn't know more specifically where C lives. Whether this is a 'quiet violation' that is intended to deceive or a maxim clash because A cannot be more specific without violating the Maxim of Quality (i.e., saying something false) depends on whether A does, in fact, know more specifically where C lives. Suppose, for example, that A knows that C does not want B dropping by unexpectedly; in this case A might choose to be purposely unclear, as in (15), in order to give the false impression of not knowing precisely where C lives.[2]

Tests for conversational implicature

Implicatures based on the CP are called **conversational implicatures**, and they differ in important ways from other types of meaning, including both semantic meaning and other types of pragmatic meaning (like conventional implicatures and presuppositions, to be discussed later). Grice proposes that a conversational implicature in general is:

- calculable
- cancelable
- nondetachable
- nonconventional
- 'not carried by what is said, but only by the saying of what is said'
- indeterminate

Not all of these are equally helpful as tests for conversational implicature (see Sadock 1978 for a great discussion of this), but let's run through them, using example (1), repeated here as (16):

(16) I'll rewrite this chapter or I'll delete it.

[2] And if you're thinking this also looks like a violation of Quantity, you're right; hold on for a discussion of newer approaches that address this issue of maxim overlap.

Here, the sentence is taken to mean 'one or the other but not both' (the **exclusive-*or*** meaning), despite the fact that the logical operator corresponding to the English word *or* (∨) means 'one or the other or both' (the **inclusive-*or*** meaning). And this sort of 'divergence in meaning' between the logical operators and their natural-language counterparts was the problem Grice undertook to solve by proposing the CP. So where does the exclusive meaning come from?

In asking that question, we're asking about **calculability**. If the hearer can't calculate the implicature, it's pointless to try to generate an implicature at all. But we can nicely calculate the exclusive meaning from a combination of the inclusive meaning and our old friend scalar implicature (see (4)). In (16), presumably if I plan to both rewrite the chapter **and** delete it, I should say so; given that I didn't, and that—based on the maxim of Quantity—you know that I should have said as much as I could, you can assume that I was not in a position to say that I would do both. The phrases *A and B* and *A or B* form a scale, with the *or* variant lower on the scale (since *A and B* entails *A or B* but not vice versa). Therefore, as with all scalar implicatures, we are licensed to infer that the uttered value holds but that no higher value holds—which is to say, in this case, that you can infer that the lower value *A or B* holds, but not the higher value *A and B*; hence, I'll do one or the other but not both. This, as shown in Horn 1972, is how we can use the CP to solve the problem of the divergence between the meaning of the logical operator ∨ and its natural-language counterpart *or*.

The second test is **cancelability**, also sometimes called **defeasibility**. This, in short, means that a conversational implicature can be immediately canceled without a sense of contradiction. For example:

(17) Today I'm going to mow the lawn or pull the weeds; in fact, I'll try to do both.

Here, the *or* in the first clause may give rise to an implicature of 'not both', as in (16), but the clause after the semicolon cancels it.

Now let's assume that I plan to both rewrite the chapter and delete it (being, I guess, a glutton for punishment). I could cancel the implicature of 'not both' by writing something like (18):

(18) I'll rewrite this chapter or I'll delete it; in fact, I plan to do both.

Now, you might find this a bit odd. After all, if I plan to do both, why not just say so from the beginning? But the cancellation in (19) is perfectly reasonable:

(19) I'll rewrite this chapter or I'll delete it—or, if the rewrite ends up being terrible, I may do both.

Entailments, on the other hand, cannot be canceled:

(20) #I'll rewrite the chapter and delete it, but I won't delete it.

Rewriting and deleting entails deleting, and that entailment—unlike an implicature—cannot be canceled. Sadock (1978) notes that just as an implicature can be canceled without contradiction, it can be **reinforced** without redundancy. Compare the reinforced implicature in (21) with the reinforced entailment in (22):

(21) I'll rewrite this chapter or I'll delete it, but not both.
(22) #I'll rewrite this chapter and delete it, and I'll delete it.

In (21) the implicature in the first clause is 'not both', and adding this explicitly does not sound redundant. In (22), on the other hand, 'I'll delete it' is entailed by the first clause, and adding this explicitly in the second clause sounds bizarrely redundant. (Note, however, that there are certain cases in which such reinforcement is possible, e.g., *I'll rewrite this chapter before I delete it, but I will delete it*; see Horn 1991.)

Cancelability is the clearest and most reliable of the tests in terms of distinguishing implicature from entailment. We'll run through the others somewhat more quickly. Conversational implicatures are **nondetachable**, which means that they can't be detached from that particular semantic meaning in that particular context; any other way of saying the same thing in the same context will give rise to the same implicature. So (23) will still convey 'not both':

(23) I'm going to do a rewrite of this chapter or I'm going to delete the thing.

Conversational implicatures are also, by definition, **nonconventional**, which simply means that they are not part of the conventional semantic meaning of the sentence. To add *but not both* as in (21) renders 'not both' part of the conventional meaning, and then it's no longer cancelable or reinforceable, as seen in (24a)–(24b), respectively:

(24) a. #I'll rewrite this chapter or I'll delete it, but not both, and maybe both.
b. #I'll rewrite this chapter or I'll delete it, but not both, and not both.

That is, once you've made the potential implicature explicit, it's part of the conventional meaning and not an implicature.

Grice's fifth test is a bit of a head-spinner. He says that the implicature is '**not carried by what is said, but only by the saying of what is said**'. What on earth does that mean? Well, it means that the implicature doesn't arise only because of the semantics of the utterance, but rather because the speaker has chosen to utter that semantic content right here, right now, in this context. Consider (25):

(25) A: I'm allergic to peanuts and wheat. I hope this cookie vendor has some without them.
B: No, I'm afraid that in every cookie you'll get peanuts or you'll get wheat.

Here, the 'not both' implicature vanishes. What this means is that it's not the semantics of 'you get X or Y' that generates the implicature of 'not both', but rather the utterance of that semantic meaning in a context that is conducive to the implicature of 'not both'. The implicature isn't carried by what is said—the semantics of the sentence—but rather by the saying of it in the given context.

Finally, the implicature is **indeterminate**. That means that even though it's calculable—that is, there's a path of reasoning that can get

you from this utterance in this context to the intended implicature—the implicature is nonetheless not a rock-solid conclusion; it's possible for a single utterance to license more than one distinct implicature, such that the implicature intended by the speaker is not the same as the inference drawn by the hearer.

The Gricean model of meaning

For Grice, then, meaning is broken down into two broad categories—natural meaning and non-natural meaning. Linguistic meaning is non-natural, and breaks down into two further broad categories—what is said and what is implicated. What is said equates to what is said semantically; what is implicated equates to pragmatics. Within the category of implicature, we get a couple of distinctions we haven't discussed yet. First, there's a distinction between **conversational** implicature—which is the kind we've been talking about thus far—and **conventional** implicature. Conversational implicature is characterized by all those properties that the 'tests for conversational implicature' test for, and they largely boil down to whether the implicature in question is conventionally attached to the linguistic expression in question. If it's not, then it must be calculable (from which it follows that it must be nondetachable), it can be canceled or reinforced, it's nonconventional by definition, it's not carried just by what is said (context matters), and it's indeterminate—and therefore, it's a conversational implicature. If, on the other hand, it **is** conventionally attached to the expression, then you don't need to calculate it, it can't be canceled (and it would be odd to reinforce it), it's conventional by definition, it is indeed carried just by what is said, and it's determinate. But you should be thinking, wait a minute—that just means its semantic meaning. What kind of implicature would have all those properties?

Welcome to a new category: **Conventional** implicature. Suppose we define semantics as truth-conditional meaning (as many linguists do). Our tests for conversational implicature, as we've seen, assume that conversational implicature is defined as nonconventional meaning. So far, so good. But what happens if there's a type of conventional meaning that isn't semantic—that is, a type of conventional meaning that isn't

part of the truth-conditions of the sentence? Well, that's precisely what we've got with conventional implicature. Remember from the end of Chapter 2 the discussion of examples like this:

(26) Now I was working against much more powerful forces, was threatened at much higher stakes, and yet my appreciation and gratitude were abundant.

And recall that the contrast associated with the word *yet* doesn't contribute anything to the truth-conditions of the sentence (flip back to the last couple of pages of Chapter 2 if you need a refresher). That is, even if there's no contrast between working against powerful forces and feeling grateful, (26) can still be considered true (assuming there's nothing else there that's false); the presence of the word *yet* in the absence of contrast doesn't in itself mean the statement is false. So the contrast in question isn't truth-conditional and therefore isn't part of the semantic meaning of (26), despite being conventionally attached to the word *yet*.

Instead, in the Gricean model, the contrast associated with the word *yet* is said to be **conventionally implicated**. It's conventional because it's impossible to use the conjunction *yet* without having the meaning of 'contrast' conveyed (and it'll therefore fail all those tests for conversational implicature), but it's an implicature because it's not part of 'what is said' truth-conditionally and hence not part of the semantics.[3] Other examples of conventional implicature include the causation associated with *therefore* and the contrast associated with *but*.

Finally, within the category of conversational implicature we have a breakdown between **generalized** and **particularized** conversational implicatures. Generalized implicatures are cases in which the implicature holds over an entire class of situations. For example, saying *Can you X?* will frequently implicate 'please X', and indicating a quantity will generally implicate 'no more than that quantity', as in (27a) and (27b), respectively:

[3] This assumes a truth-conditional semantics. See Potts 2005 for an account that does place these cases firmly within semantics.

(27) a. Can you mow the lawn this afternoon?
b. I'll give you $10 for that sweater.

In (27a), the usual implicature would be 'please mow the lawn this afternoon', and in (27b), it would be 'I won't give you more than $10 for that sweater'. But there are cases in which the generalized implicature doesn't hold:

(28) a. Can you speak Russian?
b. You need to be 21 to enter this bar.

In most situations, (28a) would not be taken as a request to speak Russian, and the speaker in (28b) wouldn't be taken to mean that 22-year-olds are forbidden to enter.

In contrast with generalized implicatures, particularized implicatures are specific to one particular utterance in one particular context. So compare the Quantity-based implicature in (27b) with that in (5), repeated here:

(29) "This is your mother," said Dorothea, who had turned to examine the group of miniatures. "It is like the tiny one you brought me; only, I should think, a better portrait. And this one opposite, who is this?"

"Her elder sister. They were, like you and your sister, the only two children of their parents, who hang above them, you see."

"The sister is pretty," said Celia, implying that she thought less favorably of Mr. Casaubon's mother.

In this situation, with two girls pointed out in a picture, to state that one of them is pretty—especially the less situationally relevant one—does indeed implicate that the unmentioned one is less pretty. But we wouldn't want to say that there's a generalized implicature to the effect that when an individual is complimented, all other salient individuals are thereby insulted. We can summarize the sort of Quantity implicature we see in (27b) with a generalization—'in general, expressing a quantity implicates that no higher quantity holds (unless higher quantities are irrelevant)'—but there's no obvious generalization that straightforwardly captures the implicature in (29); it's essentially a nonce implicature, calculated on the spot.

Thus, we can outline the Gricean model of meaning as follows:

I. Meaning
  1. natural
  2. non-natural
    a. what is said
    b. what is implicated
      i. conventionally
      ii. conversationally
        (1) generalized
        (2) particularized

Generalized conversational implicatures have a conventional aspect to them, in that they generally hold across situations and thus presumably don't need to be calculated anew each time, but they're nonetheless conversational and not conventional implicatures, because they don't hold in all contexts, they certainly **can** be calculated, and they can also be canceled or reinforced.

Implicature after Grice

Many scholars after Grice have noted that there's a certain amount of overlap in his maxims; for instance, to be relevant is to say no more than necessary, and vice versa; and to say no more than necessary is to be brief, and vice versa (and hence to be brief is to limit yourself to what's relevant, and vice versa). In light of this overlap, a number of researchers after Grice have offered alternative sets of maxims to explain the inferences hearers draw in conversation. Here I will briefly mention the three most influential.

Horn (1984, 1993) restructures the maxims as a tension between two opposing principles, which he terms Q and R:

The Q-principle: Say as much as you can, given R.
The R-principle: Say no more than you must, given Q.

As you can see, each principle builds in the acknowledgment of the opposing force and hence the tension between the two. The Q-principle corresponds to Grice's first submaxim of Quantity (say enough) and the first two submaxims of Manner (avoid obscurity and ambiguity), while the R-principle corresponds to Grice's maxim of Relation (be relevant), the second submaxim of Quantity (don't say too much), and the last two submaxims of Manner (be brief and orderly). Quality, meanwhile, is considered an umbrella maxim without which communication is impossible.

So in Horn's view, a scalar implicature is Q-based: If I say (30a), I will generally Q-implicate (30b):

(30) a. I've washed most of the windows.
b. I didn't wash all of the windows.

The Q-principle tells me to say as much as I can (given R), so if I said *most* instead of *all*, it must be the case that *all* doesn't apply. So the Q-principle licenses an inference to, in effect, 'no more than was said'.

The R-principle, on the other hand, licenses an inference to more than was said: If I say (31a), I implicate (31b):

(31) a. I chipped a tooth.
b. The tooth was my own.

It's certainly possible to chip someone else's tooth, but the default case is to chip one's own tooth, and so if I had chipped someone else's, Q would have required me to say so. In the absence of anything to indicate that it was someone else's tooth, the hearer can infer that it was my own. And by the same logic, if I say (32a), I implicate (32b):

(32) a. Gertrude was able to fix the car.
b. Gertrude fixed the car.

The R-principle tells me to say no more than I must (given Q), so even though I didn't say that Gertrude fixed the car, by saying she was able to do so I implicate that not only was she able, but she in fact did fix it.

But you'll immediately notice that these two principles license implicatures that are precisely opposite to each other: Q says 'infer that no more than this holds', while R says 'infer that something more holds'. Horn calls Q a 'lower-bounding' principle that licenses 'upper-bounding' implicatures (i.e., the speaker must say 'at least this much'—that is, as much as possible, given R—and thus implicates 'no more'), and calls R an 'upper-bounding' principle that licenses 'lower-bounding' implicatures (i.e., the speaker must say 'no more than this'—that is, no more than necessary, given Q—and thus may implicate more that was deemed not necessary to say). The Q-principle is hearer-based, since it's in the hearer's interest for the speaker to say as much as possible; it makes the hearer's job easier. And the R-principle is speaker-based, since it's in the speaker's interest to say no more than necessary; it makes the speaker's job easier.

Thus, the speaker's and hearer's opposing interests create a tension that is captured in the Q/R system. And which principle wins out in any given case determines whether the implicature will limit or extend the hearer's inference. But that in turn raises the thorny question of how we know which principle to apply in any given case. It's straightforward enough to establish in a post hoc way which one **did** apply in a particular case, by looking at which way the inference went, but unless we can use the theory to predict which principle will be used in a future instance, the theory is unfalsifiable.

Fortunately, Horn provides a way to negotiate the tension between Q and R, in what he calls the Division of Pragmatic Labor, which says in essence that an unmarked utterance licenses an R-inference to the unmarked situation, while a marked utterance licenses a Q-inference to the effect that the unmarked situation doesn't hold. So the unmarked utterances in (33a) and (34a) will license an inference to the unmarked situations in (33b) and (34b), respectively, while a relatively long or unusually phrased utterance like those in (35a) and (36a) will license an inference to a marked interpretation, as in (35b) and (36b), respectively.

(33) a. I was able to help Kris.
b. I helped Kris.

(34) a. After work I like to go home and have a drink.
b. The drink contains alcohol.

(35) a. I had the ability to help Kris.
b. I didn't necessarily help Kris.

(36) a. After work I like to go home and have a beverage.
b. The beverage doesn't necessarily contain alcohol.

Horn's theory is considered 'neo-Gricean' in that it retains Grice's essential insight that the effort to negotiate often-conflicting maxims gives rise to implicatures. Grice explicitly noted, for example, that in a scalar implicature, it's the speaker's inability to say more without violating Quality that gives rise to the 'no more than this' implicature.

Another neo-Gricean framework that is based in this same tension is that of Levinson (2000). Levinson's system is similar to Horn's except that Levinson presents three heuristics for interpreting utterances:

The Q-heuristic: What isn't said, isn't.
The I-heuristic: What is simply described is stereotypically exemplified.
The M-heuristic: A marked message indicates a marked situation.

Levinson's Q-heuristic, like Horn's Q, licenses scalar implicatures based on what the speaker has chosen not to say: If I say *I want to buy two sweaters*, I implicate that I don't want to buy three, on the grounds that what I didn't say, doesn't hold. The I-heuristic, like Horn's R, accounts for implicatures like those in (31) and (32), allowing the hearer to draw an inference from a simple utterance to a stereotypical situation.

Finally, the M-heuristic specifically addresses the form (rather than the informativeness) of an utterance. Whereas a standard scalar inference from *three* to *not four* is based on the meanings of the words *three* and *four*, not on the word *four* being any longer, more complex, or less common than the word *three* (and is therefore Q-based for Levinson), the inferences in (35) and (36) are based on the length of complexity of the expressions (and are therefore M-based for Levinson). So, for example, the phrase *had the ability to* is longer, more complex, and less common than the phrase *was able to*, and so the speaker's use of

the more marked *had the ability to* suggests a purposeful avoidance of the phrase *was able to* along with the implicature it would carry ('I did'). The markedness here isn't semantic (since *was able to* and *had the ability to* have the same semantic meaning), but rather formal—i.e., based on the form of the utterance. Similarly, the phrase *have a beverage* is more marked formally than the phrase *have a drink*, and again its use implicates that the implicature associated with the less marked option doesn't hold.

The third highly influential theory is **Relevance Theory** (Sperber and Wilson 1986). Unlike Horn and Levinson, whose theories retain Grice's tension between potentially conflicting communicative demands, Relevance Theory essentially boils all of Grice down to one overriding maxim, which corresponds roughly to Grice's Relation ('be relevant'). For Sperber and Wilson, relevance is central to human cognition, and therefore also to human communication.

It's worth noting that Grice made the same point with respect to his Cooperative Principle, i.e., that it's not just about communication but in fact it's a more general principle enjoining us to behave cooperatively. So if you ask me for a glass of beer, it would be a violation of Quantity to bring you a keg (however much you might like that) or to bring you three glasses of beer and two glasses of lemonade. Just as for Grice human cognition in general enjoins us to be maximally cooperative, for Sperber and Wilson human cognition in general enjoins us to be maximally relevant. In short, Sperber and Wilson follow Grice in taking the principles of communication to follow from principles of cognition more generally.

Sperber and Wilson offer two guiding principles:

Cognitive Principle of Relevance: Human cognition tends to be geared to the maximization of relevance. (Wilson and Sperber 2004)

Communicative Principle of Relevance: Every ostensive stimulus conveys a presumption of its own optimal relevance. (Wilson and Sperber 2004)

The first of these isn't specific to communication; the second is essentially an application of the first to communication. That is, if all human cognition is geared toward relevance, then any act of speaking can be

assumed to be relevant. Relevance itself is defined in terms of **positive cognitive effects**—cognitive changes in the way one sees the world. The assumption of relevance, combined with context, causes the hearer to search for **contextual implications,** which are essentially all the conclusions that the new utterance in combination with the context might lead the hearer to draw, including what Grice termed conversational implicatures. Contextual implications are one type of positive cognitive effect.

For Sperber and Wilson, these contextual implications are based on the context, the utterance, and the human tendency to maximize relevance. Given what the speaker has said, and given that this utterance comes with a presumption of its own optimal relevance, my job as a hearer is to figure out what the optimally relevant intended meaning is, in light of all of the usual contextual factors (who said it, when, where, etc.). I need to take the 'path of least effort' in seeking out contextual implications until my expectation of relevance has been met, at which point I can infer that I have landed at the intended meaning. For example, in (32a), when the speaker says *Gertrude was able to fix the car*, it's my job as hearer to put that together with the context (e.g., maybe I already knew that Gertrude was hoping to fix the car), and the Communicative Principle of Relevance (the intended meaning of *Gertrude was able to fix the car* is somehow optimally relevant), and put them together: If she wanted to fix it, and was able to fix it, optimal relevance suggests that she did indeed fix it.

An utterance is relevant to a hearer only when it offers contextual implications. In coming to a conclusion about a speaker's meaning, a hearer will examine possible interpretations and choose the one that gives the highest relevance, which is to say the greatest number of contextual implications—in essence, the biggest communicative bang for the buck. For Sperber and Wilson, then, an utterance like *Gertrude was able to fix the car* will, in most contexts, lead to the conclusion 'she fixed it' because, again in most contexts, her ability to fix the car is relevant only if she did in fact fix it; that is, without that contextual implication, there's no other obvious one to be drawn. Therefore, 'she fixed it' is the interpretation with the most contextual implications, and therefore the most relevance; thus, it is the likeliest interpretation.

So, like the neo-Gricean accounts, Relevance Theory involves a tension between opposing tendencies; in Relevance, it's a tension between minimizing processing effort and maximizing cognitive effect. The higher the processing cost, the lower the relevance, and the higher the cognitive payoff, the higher the relevance. You're essentially shopping at the Cognitive Effects store, with Processing Cost as the price, and looking for the best bargain. The most relevant inference will be the one that provides the greatest effect for the smallest effort, and that in turn will be the preferred interpretation.

Finally, the truth-conditional content of an utterance, including any referential, contextual, etc., information that must be filled in to render a truth-evaluable proposition, is called the **explicature** (Sperber and Wilson 1986, Carston 2002). Thus, if John Doe says (37), that statement doesn't yet represent a truth-evaluable proposition:

(37) I haven't eaten yet.

If we found (37) scribbled on a piece of paper on the sidewalk, we wouldn't have any idea whether it was true or false. In order to assign it a truth-value, we first have to know who *I* refers to, and second, we have to know what counts as *eating*. (If I say *I haven't eaten* at 10 a.m., it means something different from saying it at 10 p.m., and in neither case is it falsified if I've eaten a single potato chip a half hour earlier; see, e.g., Recanati 2004.) And finally, we need to know how long counts as *yet*; someone saying they haven't eaten yet can't mean they've never eaten in their entire life. So if we fill in who made the utterance and what counts as 'eating' and what counts as 'yet', we get something like the following:

(38) John Doe hasn't eaten breakfast yet today.

This is the explicature, and it is truth-evaluable; we can look at the world and determine whether or not it's true. (Although, of course, I'm glossing over the gradient issue of what counts as 'breakfast': Does one strip of bacon count? One egg? One of each? Or is it purely a matter of what John himself counts as breakfast? Clearly Prototype Theory can be invoked here.)

Conclusion

This chapter has breezed through a great deal of material. Grice's Cooperative Principle and its maxims and submaxims have been hugely influential in the field of pragmatics, and laid the groundwork for a great deal of work that has followed. It has obviously influenced later approaches to implicature, including the theories of the neo-Griceans and Relevance theorists, but as we'll see in the remainder of the book, it has been fundamental to the rest of pragmatic theory as well. Because pragmatics is the study of how utterance interpretation is affected by context (where one aspect of context is the hearer's assumptions about the speaker's intentions), every aspect of pragmatics asks the question 'how did we get from that utterance to this meaning?'—and that's the question Grice's work gave us some initial tools to answer. Thus, the Cooperative Principle will reverberate throughout the book as the basis for inferences in virtually every other area of pragmatics.